Biomedical Text Mining For Disease Gene Discovery

Sarah ElShal[1,2], Jesse Davis[3], and Yves Moreau[1,2]

KU LEUVEN

iMinds CONNECT.INNOVATE.CREATE

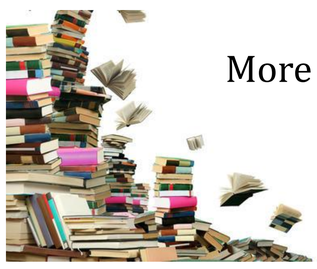
WHY: Disease Gene Discovery

Would **you** like your genome to be screened for **Alzheimer's Disease**?

What if you could get your DNA examined against the genes known to be linked to Alzheimer's Disease?
What if you realized you would get Alzheimer's Disease in a later stage of your life?

What if you could take early medications to limit depression and irritability as two early symptoms?

What kind of life style would you choose?



More than **23 million citations** for biomedical literature.

PubMed

Linking genes to diseases is one challenge that can be approached differently.
- Some researchers analyze the whole human genome to find genes that are similar to the known disease –causing genes
- We text-mine the literature in PubMed to find potential links between diseases and genes.

WHAT: Google-like Tool

Beegle

Clinical Experts

Alzheimer's Disease

PSEN1
Presenilin 1
The **PSEN1**, p.E318G Variant Increases the Risk of **Alzheimer's Disease** in APOE-ε4 Carriers.

SORL1
Sortilin related receptor
An updated Meta-Analysis of the Association between **SORL1** variants and the Risk for Sporadic **Alzheimer's Disease.**

PSEN2
Presenilin 2
Several mutations of the genes APP, PSEN1 and **PSEN2** are described. These cause around half of all cases of the rare early onset autosomal dominant form of **Alzheimer's disease**

. . . .

Given: Any free text input (disease/disorder/biological process...)
Return: An ordered list of the genes most linked to the given input.
→ Hence the **clinical expert** can further experiment the top genes in this list and validate which of them can be defined as disease-causing.

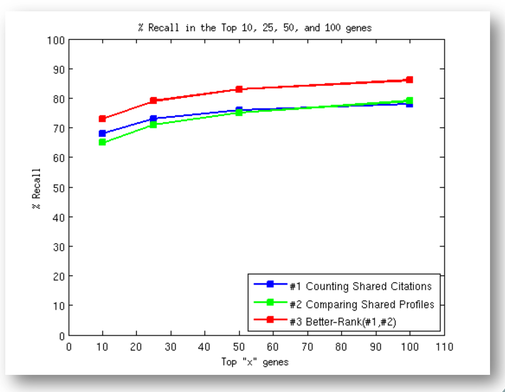
Alzheimer's Disease is just an example Beegle is designed to work for any free-text input

RESULTS:

86% Recall in top 100

We evaluate our tool based on the disease-gene associations in the **OMIM catalog**. Our best performance is reflected in achieving a recall of **73% and 86% in the top returned 10 and 100** genes respectively.

These results show the potential for text mining to discover links between diseases and genes in the biomedical literature. Since a high percentage of the top ranking links is already experimentally-validated, we can highlight the other percentage as potential candidates for further validation.



HOW: Text Mining PubMed

For each gene recorded in "Entrez Gene", we record the list of PubMed abstracts linked to it as indicated on GeneRIF. Then for each disease query, we record its list of abstracts as retuned by the online query system of PubMed. To measure the strength of the association between a disease and a given gene we investigate **three approaches**.

#1 Counting shared Citations

Alzheimer's Disease

N
- Caspase-6 is an effector caspase that has not been investigated thoroughly despite the fact that Caspase-6 is strongly activated in **Alzheimer disease** brains... PMID: 24265764
- . . . .

X
- Less than 5% of all **Alzheimer's disease** cases are familial in nature, i.e. caused by mutations **in APP, PSEN1 or PSEN2**... PMID: 24101602
- Presenilin 1 **(PSEN1) gene mutations** deterministic for **Alzheimer's disease (AD)** are associated with marked heterogeneity in clinical phenotype... PMID: 23948899
- . . . .

X

PSEN1

K
- Upon **PSEN1**/2 re-introduction, this active epigenetic state was replaced by a MeCP2-containing repressive state and reduced Neurotrypsin expression.. PMID: 24145027
- . . . .

	query	$\overline{query}$	
gene	**X**		**K**
$\overline{gene}$			
	N		**M**

$$score = p - value = \frac{\binom{K}{X}\binom{M-K}{N-X}}{\binom{M}{N}}$$

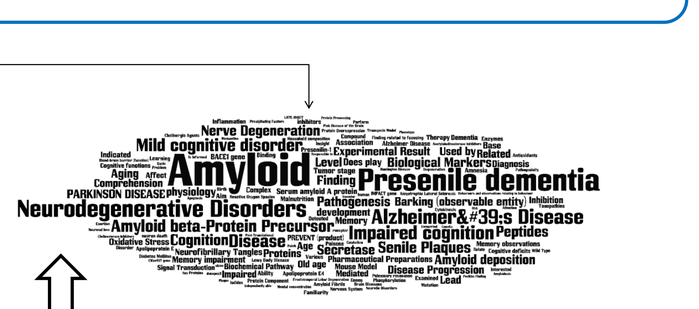
M=#background_citations

In the first approach, we rely on the **clear association** between the disease and the gene as reported in the literature. The higher the count of abstracts that mention both of them, the stronger the association is. We use **Fisher's exact test** to evaluate such association.

#2 Comparing Text Profiles

Alzheimer's Disease

Alzheimer's Disease	Amyloid	Presenile dementia	. . . .
0.034075	0.074095	0.051059	. . . .



Alzheimer 'sDisease	Amyloid	Presenile dementia	. . . .
0.041778	0.057728	0.021259	. . . .

PSEN1

$$score = \frac{(tf \times \log(idf)_{query} \bullet tf \times \log(idf)_{gene})}{\| tf \times \log(idf)_{query} \| \| tf \times \log(idf)_{gene} \|}$$

In the second appraoch we go a further step and try to discover **hidden associations** between the disease and the given gene. Given the list of abstracts linked to a disease/gene, we generate a term profile that defines this disease/gene. The higher the count of shared terms between the disease and the gene profiles, the stronger the association is. We use the **cosine similarity measure** to evaluate such association.

#3 Better_rank(#1, #2)

Alzheimer's Disease Genes	#1	#2	#3
SORL1	**5**	51	**7**
PSEN2	**7**	8	**11**
PSEN1	**3**	5	**3**

The third appraoch is a **combined approach** where we consider the two association signals coming from the first and the second approaches.
For a given gene, we select the stronger signal to account for its association to the given disease.

[1]KU Leuven, Department of Electrical Engineering (ESAT) STADIUS, 3001 Leuven, *Belgium*
[2]iMinds Future Health Department, Leuven, *Belgium*
[3]KU Leuven, Department of Computer Science (DTAI), 3001 Leuven, *Belgium*

sarah.elshal@esat.kuleuven.be
jesse.davis@cs.kuleuven.be
yves.moreau@esat.kuleuven.be

SISTA

DTAI Declaratieve Talen en Artificiële Intelligentie